AFRL-AFOSR-JP-TR-2017-0041

Algorithms that defy the gravity of learning curve

Kai Ming Ting
FEDERATION UNIVERSITY AUSTRALIA

04/28/2017
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ IOA
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* 25-05-2017	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* 30 Apr 2015 to 29 Apr 2017

4. TITLE AND SUBTITLE Algorithms that defy the gravity of learning curve	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER FA2386-15-1-4009
	5c. PROGRAM ELEMENT NUMBER 61102F
6. AUTHOR(S) Kai Ming Ting	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) FEDERATION UNIVERSITY AUSTRALIA UNIVERSITY DR MOUNT HELEN, 3350 AU	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AOARD UNIT 45002 APO AP 96338-5002	**10. SPONSOR/MONITOR'S ACRONYM(S)** AFRL/AFOSR IOA
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-AFOSR-JP-TR-2017-0041

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Conventional wisdom posits that the learning behaviour of all data mining algorithms follows a typical learning curve, where more data is expected to produce better performing models. We call this behaviour the gravity of learning curve which all algorithms are assumed to comply. This project provides theoretical analysis and empirical evidence for the first time that nearest neighbour anomaly detectors defy the gravity of learning curve, i.e., these gravity defiant algorithms can learn a better performing model using a small training set than that using a large training set. The knowledge we uncovered enables algorithms to be utilized in a new way to meet the challenges of big data without ever-increasing demands for big data infrastructures.

This project has spent a significant amount of time perfecting the theory and conducting a rigorous empirical evaluation. As a result, the insight gained is much better than we anticipated. The outcome is a major publication in Machine Learning Journal, published in early 2017. In addition, during this project period, four papers from two previous AOARD supported projects have been published. These include a major work on mass-based dissimilarity which was published in The ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2016. This work has informed one of the investigations in this project.

15. SUBJECT TERMS
Data Mining

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			KNOPP, JEREMY
Unclassified	Unclassified	Unclassified	SAR	29	**19b. TELEPHONE NUMBER** *(Include area code)* 315-227-7006

Final Report for AOARD Grant FA2386-15-1-4009

"Algorithms that defy the gravity of learning curve"

Date: April 28, 2017

Name of Principal Investigator: Kai Ming Ting

E-mail: kaiming.ting@federation.edu.au
Institution: Federation University
Mailing Address: PO Box 3191 Gippsland Mail Centre VIC 3841, Australia
Phone: +61 3 512 26241

Period of Performance: 30 April, 2015 - 29 April, 2017

Abstract

Conventional wisdom posits that the learning behaviour of all data mining algorithms follows a typical learning curve, where more data is expected to produce better performing models. We call this behaviour the gravity of learning curve which all algorithms are assumed to comply. This project provides theoretical analysis and empirical evidence for the first time that nearest neighbour anomaly detectors defy the gravity of learning curve, i.e., these gravity defiant algorithms can learn a better performing model using a small training set than that using a large training set. The knowledge we uncovered enables algorithms to be utilized in a new way to meet the challenges of big data without ever-increasing demands for big data infrastructures.

This project has spent a significant amount of time perfecting the theory and conducting a rigorous empirical evaluation. As a result, the insight gained is much better than we anticipated. The outcome is a major publication in Machine Learning Journal, published in early 2017. In addition, during this project period, four papers from two previous AOARD supported projects have been published. These include a major work on mass-based dissimilarity which was published in The ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2016. This work has informed one of the investigations in this project.

1 Introduction

In the age of big data, the revelation and the new knowledge about the gravity defiant algorithms created in this project have two impacts. First, the capacity provided by big data infrastructures would be overkill because the gravity defiant algorithms that produce good performing models using small data sets can be executed easily in existing computing infrastructures. The resources, that would otherwise be used to build big data infrastructures, can now be employed for other projects that yield a better return in investment. Second, this project will open a whole new direction of research into different types of gravity defiant algorithms that are readily applicable to mine big data without ever-increasing demands of big data infrastructures, as the data size increases over time. This project focuses on nearest neighbour-based anomaly detectors because they have been shown to be one of the most effective class of anomaly detectors reported in the literature.

Although there are a few works in the literature which provide an indication that some algorithms may defy the gravity of learning curve, no concrete evidence has been provided, let alone a theoretical analysis. This is the first theoretical work, as far as we know, which investigates algorithms that defy the gravity of learning curve. It also provides concrete empirical evidence that there are gravity-defiant algorithms which produce good performing models with small training sets; and models trained with large data sizes perform worse.

This project aims to

1. **Provide evidence that there are algorithms which defy the gravity of learning curve.**
2. **Offer an insight into the behaviour of gravity defiant algorithms through theoretical analyses.**
3. **Identify the key differences between the gravity defiant ensemble approach and the conventional ensemble approach.**
4. **Reveal the effect of new mass-based similarity measures on conventional algorithms and new algorithms in terms of learning curve.**

This report provides a summary of the theoretical analyses in Section 2, the results and discussion in Section 3, and the conclusions in Section 4. To be concise, a lot of details have been omitted in this report. A detailed account of these materials can be found in the attached Paper [1].

2 Theoretical Analyses

This project makes the following contributions:

1. Provide a theoretical analysis of nearest neighbour-based anomaly detection algorithms which reveals that their behaviours defy the gravity of learning curve. This is the first analysis in machine learning research on learning curve behaviour that is based on computational geometry, as far as we know.
2. The theoretical analysis provides an insight into the behaviour of the nearest neighbour anomaly detector. In sharp contrast to the conventional wisdom: more data the better, the analysis reveals that sample size has three impacts which have not been considered by the conventional wisdom. First, increasing sample size increases the likelihood of anomaly contamination in the sample; and any inclusion of anomalies in the sample increases the false negative rate, thus, lowers the AUC. Second, the optimal sample size depends on the data distribution. As long as the data distribution is not sufficiently represented by the current sample, increasing the sample

size will improve AUC. The optimal size is the number of instances best represents the geometry of normal instances and anomalies; this gives the optimal separation between normal instances and anomalies, encapsulated as the average nearest neighbour distance to anomalies. Third, increasing the sample size decreases the average nearest neighbour distance to anomalies. Increasing beyond the optimal sample size reduces the separation between normal instances and anomalies smaller than the optimal. This leads to the decreased AUC and gives rise to the gravity-defiant behaviour.

The details of the characterisation of the accuracy of nearest neighbour anomaly detector based on computational geometry is provided in Sections 3, 4 and 5 in Paper [1]. A simplified version of the theoretical analyses is extracted in the following five subsections.

2.1 Modeling $\mathcal{X}_N$ and $\mathcal{X}$ based on computational geometry

Let $(\mathcal{M}, m)$ be a metric space, where $\mathcal{M}$ is a d dimensional space and m is a distance measure in $\mathcal{M}$. Let $\mathcal{X}$ be a d dimensional open subset of $\mathcal{M}$. $\mathcal{X}$ is split into a subset of normal instances $\mathcal{X}_N$ and a subset of anomalies $\mathcal{X}_A = \mathcal{X} \backslash \mathcal{X}_N$ by an oracle. Assume that each of $\mathcal{X}$, $\mathcal{X}_N$ and $\mathcal{X}_A$ can be partitioned into a finite number of convex d dimensional subsets.

Let D_N and D_A (in D) be sets of instances belonging to $\mathcal{X}_N$ and $\mathcal{X}_A$, respectively, *i.e.*, $D_N = \{x \in D | x \in \mathcal{X}_N\}$ and $D_A = \{x \in D | x \in \mathcal{X}_A\}$. In anomaly detection, we assume that the size of D_A is substantially smaller than D_N, *i.e.*, $|D_N| \gg |D_A|$.

We employ a nearest neighbour anomaly detector (1NN) using $\mathcal{D}$ in this analysis; and it uses an anomaly score for any $x \in D$ defined by x's nearest neighbour distance in $\mathcal{D} \subset D$ as follows:

$$q(x; \mathcal{D}) = \min_{y \in \mathcal{D}} m(x, y). \tag{1}$$

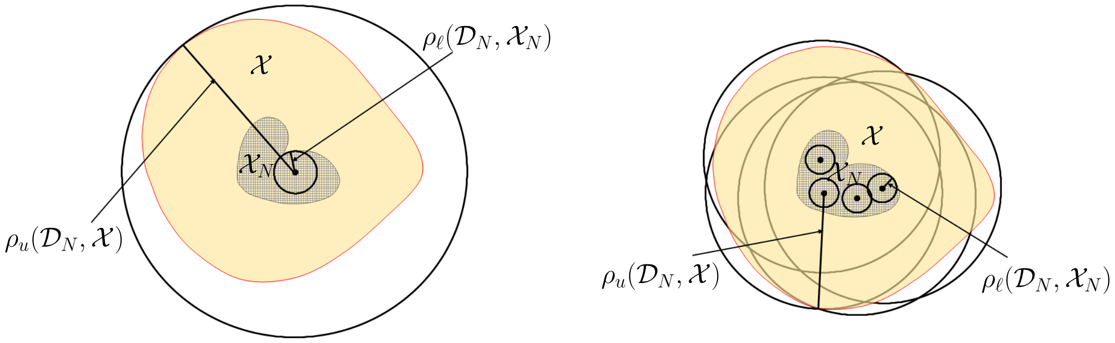
Here, we model $\mathcal{X}_N$ and $\mathcal{X}$ in $\mathcal{M}$ using computational geometry in relation to the anomaly score $q(x; \mathcal{D}_N)$.

Let the set of all points satisfying $q(x; \mathcal{D}_N) \leq r$ in $\mathcal{M}$ be a union of balls $B_d(y, r)$ for all $y \in \mathcal{D}_N$, where $B_d(y, r)$ is a d dimensional ball centered at y with radius r.

$\mathcal{X}_N$ and $\mathcal{X}$ can now be modelled using these balls which have two critical radii, *i.e.*, the inradius of $\mathcal{X}_N$: $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$; and the covering radius of $\mathcal{X}$: $\rho_u(\mathcal{D}_N, \mathcal{X})$, formally defined as follows:

$$\begin{aligned}\rho_\ell(\mathcal{D}_N,\mathcal{X}_N) &= \sup\arg_r\left[\bigcup_{y\in\mathcal{D}_N} B_d(y,r)\subseteq\mathcal{X}_N\right]\\ &= \sup\arg_r\left[\{x\in\mathcal{M}|q(x;\mathcal{D}_N)\le r\}\subseteq\mathcal{X}_N\right], \text{ and}\\ \rho_u(\mathcal{D}_N,\mathcal{X}) &= \inf\arg_r\left[\bigcup_{y\in\mathcal{D}_N} B_d(y,r)\supseteq\mathcal{X}\right]\\ &= \inf\arg_r\left[\{x\in\mathcal{M}|q(x;\mathcal{D}_N)\le r\}\supseteq\mathcal{X}\right].\end{aligned}$$

Figure 1 shows two examples of $\mathcal{X}_N$ and $\mathcal{X}$ being modelled using balls having the two radii.


$\mathcal{X}$
$\rho_\ell(\mathcal{D}_N,\mathcal{X}_N)$
$\mathcal{X}_N$
$\rho_u(\mathcal{D}_N,\mathcal{X})$
$\mathcal{X}$
$\mathcal{X}_N$
$\rho_u(\mathcal{D}_N,\mathcal{X})$
$\rho_\ell(\mathcal{D}_N,\mathcal{X}_N)$

(a) $\mathcal{D}_N$ has one instance.

(b) $\mathcal{D}_N$ has four instances.

Figure 1: Examples of $\rho_\ell(\mathcal{D}_N,\mathcal{X}_N)$ and $\rho_u(\mathcal{D}_N,\mathcal{X})$, where $\mathcal{X}=\mathcal{X}_N\cup\mathcal{X}_A$, and $\mathcal{D}_N$ is represented by points in $\mathcal{X}_N$.

2.2 Characterisation of the accuracy of nearest neighbour anomaly detector based on computational geometry

The simplified Theorem for the accuracy of nearest neighbour anomaly detector, expressed as Area Under ROC Curve (AUC) is given below.

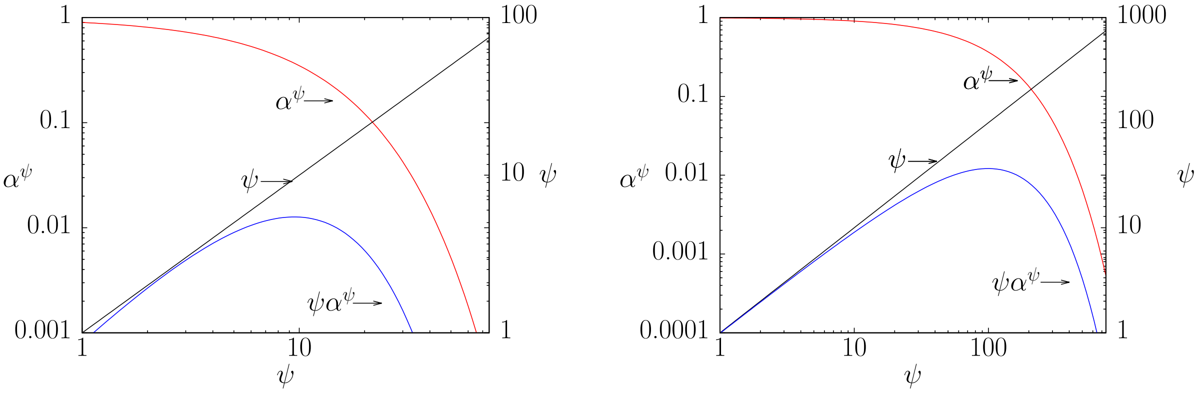
Theorem 1 *The expectation of* $\langle AUC \rangle_h$ *over the distribution of* h $(P(h))$*, i.e.,* $\langle AUC \rangle = \sum_{h=0}^{\psi} P(h) \langle AUC \rangle_h$*, has the following upper bound:*

$$\langle AUC \rangle < C\psi\alpha^{\psi}\rho_{\delta}^{b}(\psi) \tag{2}$$

where α is the proportion of normal instances estimated as $\alpha = \frac{|\mathcal{D}_N|}{|\mathcal{D}|} = \frac{h}{\psi}$.
In plain language, the three factors can be interpreted as follows: $1 - \alpha^{\psi}$ reflects the likelihood of anomaly contamination in the sample $\mathcal{D}$; ψ represents the number of balls used to represent the geometry of normal instances and anomalies; $\rho_{\delta}^{b}(\psi) = \left\langle \rho_u(\mathcal{X})^b \right\rangle_{\psi} - \left\langle \rho_{\ell}(\mathcal{X}_N)^b \right\rangle_{\psi}$ signifies the separation between anomalies and normal instances, represented by ψ balls; and C is a constant.

2.3 Gravity-defiant behaviour

As revealed in Theorem 1, the upper bound of AUC has two critical terms ψ and α^{ψ} which are monotonic functions changing in opposite directions, *i.e.*, as ψ increases, α^{ψ} decreases. Therefore, the AUC bounded by $\psi\alpha^{\psi}$ is expected to reach the optimal at some finite and positive ψ_{opt}; and the anomaly detector will perform worse if the sample size used is larger than ψ_{opt}, *i.e.*, the gravity-defiant behaviour.



(a) $\alpha = 0.9$: $\psi\alpha^{\psi}$ has $\psi_{opt} = 10$.

(b) $\alpha = 0.99$: $\psi\alpha^{\psi}$ has $\psi_{opt} = 100$.

Figure 2: The $\psi\alpha^{\psi}$ curves as a function of ψ with (a) $\alpha = 0.9$ and (b) $\alpha = 0.99$. The left and right y-axes are for α^{ψ} and ψ curves, respectively. Note that the y-axis scale of the $\psi\alpha^{\psi}$ curves is not shown.

Figure 2 shows two examples of the gravity-defiant behaviour, as a result of the upper bound. They are represented by the $\psi\alpha^{\psi}$ curves for $\alpha = 0.9$ and 0.99. This shows that the anomaly detector using anomaly score $q(x;\mathcal{D})$ has the gravity-defiant behaviour, where $\psi_{opt} < \psi_{max}$; and ψ_{max} is either the size of the given dataset or the largest sample size that can be employed. The gravity-defiant behaviour is in contrary to the conventional wisdom.

In addition, $\rho_{\delta}^{b}(\psi)$ is positive, smooth and anti-monotonic over the change of ψ (see Section 3.5 in Paper [1] for details).

In summary, both α^{ψ} and $\rho_{\delta}^{b}(\psi)$ decrease as ψ increases and wield the similar influence on the nearest neighbour anomaly detector to exhibit the gravity-defiant behaviour.

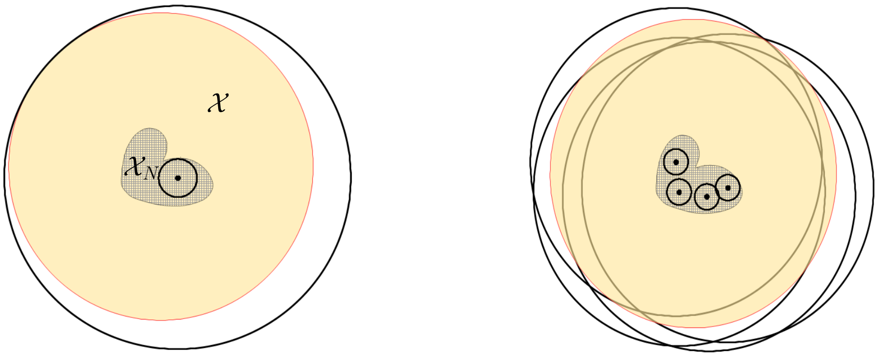
2.4 Analysis of factors which influence the AUC of the nearest neighbour anomaly detector

Here we further analyse three factors which influence the AUC of the nearest neighbour anomaly detector:

a. The proportion of normal instances, α : As shown in the last section, the nearest neighbour anomaly detector is expected to improve its AUC by the rate α^{ψ} as the proportion of normal instances increases. The change in α does not affect the other two factors, if the change does not affect the geometry of $\mathcal{X}_N$ and $\mathcal{X}$. In addition to the effect on the magnitude of AUC, Figure 2 also shows that ψ_{opt} becomes larger as α increases. This is because α^{ψ} increases as α increases.

b. The difference between the covering radius of $\mathcal{X}$ and the inradius of $\mathcal{X}_N$, *i.e.*, $\rho_{\delta}^{b}(\psi) = \left\langle \rho_u(\mathcal{X})^b \right\rangle_{\psi} - \left\langle \rho_{\ell}(\mathcal{X}_N)^b \right\rangle_{\psi}$: This factor depends on the geometry of normal clusters as well as anomalies, and influences the AUC in the following scenarios:

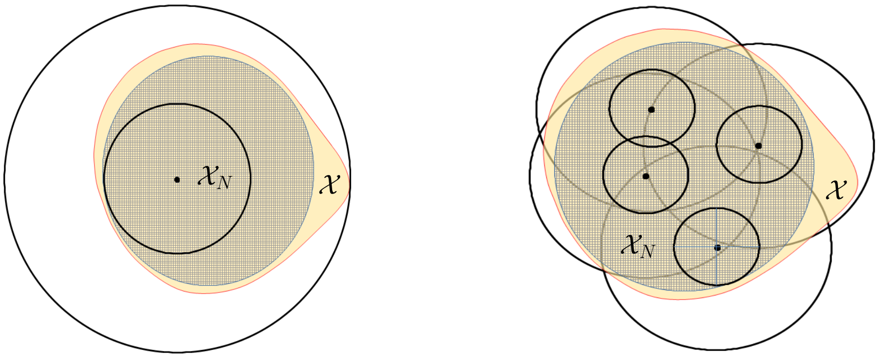
   1. $\mathcal{X}_A$ becomes bigger. The change in $\mathcal{X}_A$ with fixed $\mathcal{X}_N$ directly affects $\mathcal{X}$. Examples of this change from Figure 1 are shown in Figure 3. The enlarged $\mathcal{X}_A$, thus the enlarged $\mathcal{X}$, leads to larger $\rho_{\delta}^{b}(\psi)$ and higher AUC because the expected $\rho_u(\mathcal{D}_N, \mathcal{X})$ gets larger while the expected $\rho_{\ell}(\mathcal{D}_N, \mathcal{X}_N)$ is fixed for a given ψ.
   2. $\mathcal{X}_N$ becomes bigger. The change in $\mathcal{X}_N$ with fixed $\mathcal{X}$ affects both the expected $\rho_{\ell}(\mathcal{D}_N, \mathcal{X}_N)$ and the expected $\rho_u(\mathcal{D}_N, \mathcal{X})$. Examples of this change from Figure 1 are depicted in Figure 4. The enlarged $\mathcal{X}_N$ leads to smaller $\rho_{\delta}^{b}(\psi)$ and



(a) $\mathcal{D}_N$ has one instance.

(b) $\mathcal{D}_N$ has four instances.

Figure 3: $\mathcal{X}_A$ is larger than that in Figure 1. Examples of $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ and $\rho_u(\mathcal{D}_N, \mathcal{X})$.



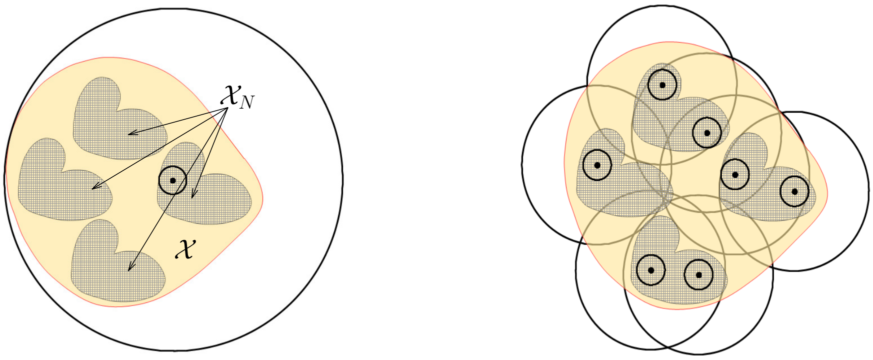
(a) $\mathcal{D}_N$ has one instance.

(b) $\mathcal{D}_N$ has four instances.

Figure 4: $\mathcal{X}_N$ and $\mathcal{X}$ have approximately the same size. Examples of $\rho_u(\mathcal{D}_N, \mathcal{X})$ and $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$.

thus lower AUC because the difference between the expected $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ and the expected $\rho_u(\mathcal{D}_N, \mathcal{X})$ gets smaller—a result of the increased $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ and the decreased $\rho_u(\mathcal{D}_N, \mathcal{X})$ for a given ψ.

3. Number of clusters in $\mathcal{X}_N$ increases. If $\mathcal{X}_N$ consists of multiple well-separated clusters as shown in Figure 5, then $\left\langle \rho_\ell(\mathcal{X}_N)^b \right\rangle_\psi$ is determined by the minimum of $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ of single clusters, regardless of the total volume or the number of clusters in $\mathcal{X}_N$. This is despite the fact that the total volume of $\mathcal{X}_N$ has increased from that in Figure 1. The expected $\rho_u(\mathcal{D}_N, \mathcal{X})$ in Figure 5 is less than that in Figure 1 because the clusters are scattered in $\mathcal{X}$. With fixed ψ, the AUC is expected to decrease in Figure 5 in comparison with that in



(a) $\mathcal{D}_N$ has one instance.

(b) $\mathcal{D}_N$ has seven instances.

Figure 5: $\mathcal{X}_N$ consists of multiple clusters. Examples of $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ and $\rho_u(\mathcal{D}_N, \mathcal{X})$.

Figure 1 because of the decreased $\left\langle \rho_u(\mathcal{X})^b \right\rangle_\psi$, which is obvious in the change from Figure 1(b) to Figure 5(b).

Anomalies' nearest neighbour distances. As indicated in b1) and b2), enlarging $\mathcal{X}_N$ has the same effect of shrinking $\mathcal{X}$ in decreasing AUC. It is instructive to note that either of these two changes effectively reduces the anomalies' nearest neighbour distances because the area occupied by $\mathcal{X}_A$ decreases. This can be seen from $\rho_\delta^b(\psi) = \left\langle \rho_u(\mathcal{X})^b \right\rangle_\psi - \left\langle \rho_\ell(\mathcal{X}_N)^b \right\rangle_\psi$, where $\rho_\delta^b(\psi)$ changes in the same direction of the expected $\rho_u(\mathcal{D}_N, \mathcal{X})$ or in the opposite direction of the expected $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$. The nearest neighbour distance of anomaly[1], which can be measured easily, is a proxy to $\rho_\delta^b(\psi)$.

In a nutshell, **any changes in $\mathcal{X}$ and $\mathcal{X}_N$ that matter—which finally vary AUC—are manifested as changes in the anomalies' nearest neighbour distances (Δ_A)**. The AUC, $\rho_\delta^b(\psi)$ and Δ_A change in the same direction.

Note that $\rho_\delta^b(\psi)$ has the same effect of α^ψ to shift ψ_{opt}. But the influence of $\rho_\delta^b(\psi)$ is more difficult to predict because it depends on the rate of decrease between the covering radius of $\mathcal{X}$ and inradius of $\mathcal{X}_N$ which in turn depend on the geometry of $\mathcal{X}$ and $\mathcal{X}_N$; and it is hard to measure in practice too.

[1] A more accurate proxy is the distance between anomaly and its nearest normal instance. In the unsupervised learning context, this distance cannot be measured easily. We will see in the experiment section that the nearest neighbour distance of anomaly is a good proxy to $\rho_\delta^b(\psi)$, even in a dataset with clustered anomalies—a factor not considered in the analysis.

c. The sample size (ψ) used by the anomaly detector : The optimal sample size is the number of instances best represents the geometry of normal instances and anomalies ($\mathcal{X}_N$ and $\mathcal{X}$). The sample size also affects two other factors, *i.e.*, as ψ increases, both α^ψ and $\rho^b_\delta(\psi)$ decrease. The direction of the change in AUC depends on the interaction between ψ and $\alpha^\psi \rho^b_\delta(\psi)$ which change in opposite directions. In general, as ψ increases from a small value, AUC improves until it reaches the optimal. Further increase from ψ_{opt} degrades AUC which gives rise to the gravity-defiant behaviour. Note that the change in ψ does not alter the data characteristics (*i.e.*, α, $\mathcal{X}_N$ and $\mathcal{X}$).

Changes in the first two factors, which affect the data characteristics, are summarised in Table 1. Each effect shown is a result of an isolated factor.

Table 1: Changes in AUC and ψ_{opt} as one data characteristic (α, $\rho_u(\mathcal{X})$ or $\rho_\ell(\mathcal{X}_N)$) changes. Δ_A is the nearest neighbour distances of anomalies.

Change in one data characteristic	$\rho_u(\mathcal{X})$	$\rho_\ell(\mathcal{X}_N)$	Δ_A	AUC	ψ_{opt}
a) α increases	=	=	=	⇑	⇑
b1) $\mathcal{X}_A$ becomes bigger	⇑	=	⇑	⇑	*
b2) $\mathcal{X}_N$ becomes bigger	⇓	⇑	⇓	⇓	*
b3) Number of clusters in $\mathcal{X}_N$ increases	⇓	=	⇓	⇓	*

* The direction of ψ_{opt} depends on the geometry of $\mathcal{X}$ and $\mathcal{X}_N$.

While one can expect the nearest neighbour anomaly detector to exhibit gravity-defiant learning curves, there are two scenarios in which only half of the curve can be observed.

- First half of the curve: For a dataset which requires large ψ_{opt}, the dataset size needs to be very large in order to observe the gravity-defiant behaviour. In the case that the data collected is not large enough, ψ_{opt} may not be achievable in practice.
- Second half of the curve: This is observed on a dataset which requires small ψ_{opt} e.g., $\psi_{opt} = 1$.

The above impacts are due to the change in sample size which, by itself alone, does not alter anomaly contamination rate or geometry of normal instances and anomalies in the given dataset (described in (a) and (b) above). Any change in geometrical data characteristics, which affects the AUC, manifests as a change in nearest neighbour distance such that the AUC and anomalies' nearest neighbour distances change in the same direction. Because nearest neighbour distance can be measured easily and other indicators of detection accuracy are difficult to measure, it provides a unique useful practical tool to detect change in domains where any such changes are critical in their change detection operations, e.g., in data streams.

2.5 Does the theoretical result apply to other nearest neighbour-based anomaly detectors?

The above theoretical analyses are based on the simplest nearest neighbour (1NN) anomaly detector with a small sample. We believe that this result applies to other nearest neighbour-based anomaly detectors too, although a direct analysis is not straightforward.

We provide a summary of our reasoning as to why the theoretical result can be applied to three nearest neighbour-based anomaly detectors, *i.e.*, an ensemble of nearest neighbours, a recent nearest neighbour-based ensemble method called iNNE [7], and k-nearest neighbour (kNN). The details are available in Section 5 of Paper [1].

Because aNNE is an ensemble of a variant of 1NN, it can be expected to behave similarly as 1NN; but it has a lower variance and the size of the variance reduction is proportional to the ensemble size. Thus, the result of the theoretical analyses applies directly to aNNE.

The analyses on kNN and iNNE are not a straightforward extension of the analysis on aNNE, and the optimal k for kNN depends on data size. Given that they all based on the basic operation: nearest neighbour, we can expect iNNE, aNNE and kNN to have the same behaviour in terms of learning curve.

In a nutshell, all three algorithms, aNNE, kNN and iNNE, can be expected to have the gravity-defiant behaviour. However, at what sample size (ψ_{opt}) each will arrive at its optimal detection accuracy is of great importance in choosing the algorithm to use in practice.

3 Results and Discussion

This section presents empirical evidence of the gravity-defiant behaviour using three nearest neighbour-based anomaly detectors, aNNE, kNN and iNNE, in the unsupervised learning context.

This project uncovers two features of nearest neighbour anomaly detectors:

a. Some nearest neighbour anomaly detector can achieve high detection accuracy with a significantly smaller sample size than others.

b. Any change in geometrical data characteristics, which affects the detection error, manifests as a change in nearest neighbour distance such that the detection error and anomalies' nearest neighbour distances change in opposite directions. Because nearest neighbour distance can be measured easily and other indicators of detection accuracy are difficult to measure, it provides a unique useful practical tool to detect change in domains where any such changes are critical in their change detection operations, e.g., in data streams. Note that the change in sample size does not alter the geometrical data characteristics discussed here.

3.1 Experimental Methodology

Algorithms used in the experiments are aNNE, iNNE and kNN (where the anomaly score of kNN is computed from the average distance of k nearest neighbours).
The experiments are designed to:

1. Verify that each of aNNE, iNNE and kNN has the gravity-defiant behaviour.

2. Compare ψ_{opt} of these algorithms.

3. Attest the effect of each of the three factors on the detection accuracy, as revealed by the theoretical analyses.

In this report, we show the key results for 1 and 2 only. The detail results for all experiments are provided in Sections 7.2 - 7.5 and Appendices in Paper [1].
The performance measure is anomaly detection error, measured as 1 - AUC, where AUC is the area under the receiver operating characteristics curve which measures the 'goodness' of the ranking result. Error = 0 if an anomaly detector ranks all anomalies at the top; and error = 1 if all anomalies are ranked at the bottom; a random ranker will produce error = 0.5.
A learning curve is produced for each anomaly detector on each dataset.

Table 2: Datasets used in the experiments.

Data	Size n	d	anomaly class
CoverType	286048	10	class 4 (0.9%) vs. class 2
Mulcross	262144	4	1% anomalies
Smtp	95156	3	attack (0.03%)
U2R	60821	34	attack (0.37%)
P53Mutant	31159	5408	active (0.5%) vs. inactive
Mammograhpy	11183	6	class1 (2.32%)
Har	4728	561	sitting, standing & laying (1.2%)
ALOI	100000	64	0.553% anomalies with 900 normal clusters

The algorithms aNNE and iNNE have two parameters: sample size ψ and ensemble size t. To produce a learning curve, the training data is constructed using a sample of size $t\psi$ where $t = 100$ and ψ is 1, 2, 5, 10, 20, 35, 50, 75, 100, 150, 200, 500 and 1000 for each point on the curve. The parameter k in kNN was set as $k = \lfloor\sqrt{n}\rfloor$ where n is the number of training instances ($n = t\psi$). Note that the minimum ψ setting is 1 for aNNE and kNN; but because each hypersphere iNNE built derives its size from the hypersphere's center to its nearest neighbour, it requires a minimum of two instances in each sample.

For an ensemble of t models, the total number of training instances employed is $t\psi$. To train a single model such as kNN, a training set of $t\psi$ instances is used in order to ensure that a fair comparison is made between an ensemble and a single model.

The Euclidean distance is used in all three algorithms.

A total of eight datasets are used in the experiment. Six datasets are from the UCI Machine Learning Repository, one dataset is produced from the Mulcross data generator[2], and the ALOI dataset is from the MultiView dataset collection[3]. They are chosen because they represent different data characteristics of data size, number of dimensions, and proportion of normal instances and anomalies. Each dataset is normalised using the min-max normalisation. The data characteristics of these datasets are given in Table 2.

The ALOI dataset has $C = 900$ normal clusters and 100 anomaly clusters, where each anomaly cluster has between 1 and 10 instances.

In every experiment, each dataset is randomly split into two equal-size stratified subsets, where one is used for training and the other for testing. For example, in each trial, the

[2] http://lib.stat.cmu.edu/jasasoftware/rocke

[3] http://elki.dbs.ifi.lmu.de/wiki/DataSets/MultiView. Accessed: 11, November 2014.

CoverType dataset is randomly split into two subsets, each has 143024 instances. The subset, which is used to produce training instances, is sampled without replacement to obtain the required t samples, each having ψ instances. As $t = 100$ and the maximum ψ is 1000, the maximum training instances employed is 100000. For datasets which have less than 200000 instances, the sampling without replacement process is restarted with the same subset when the instances have run out.

The result in each dataset is obtained from an average over 20 trials. Each trial employs a training set to train an anomaly detector and its detection performance is measured using a testing set.

3.2 Gravity-defiant behaviour

We investigate whether iNNE, aNNE and kNN have the gravity-defiant behaviour using the eight datasets in this section. The learning curves for iNNE, aNNE and kNN on each dataset are shown in Figure 6.

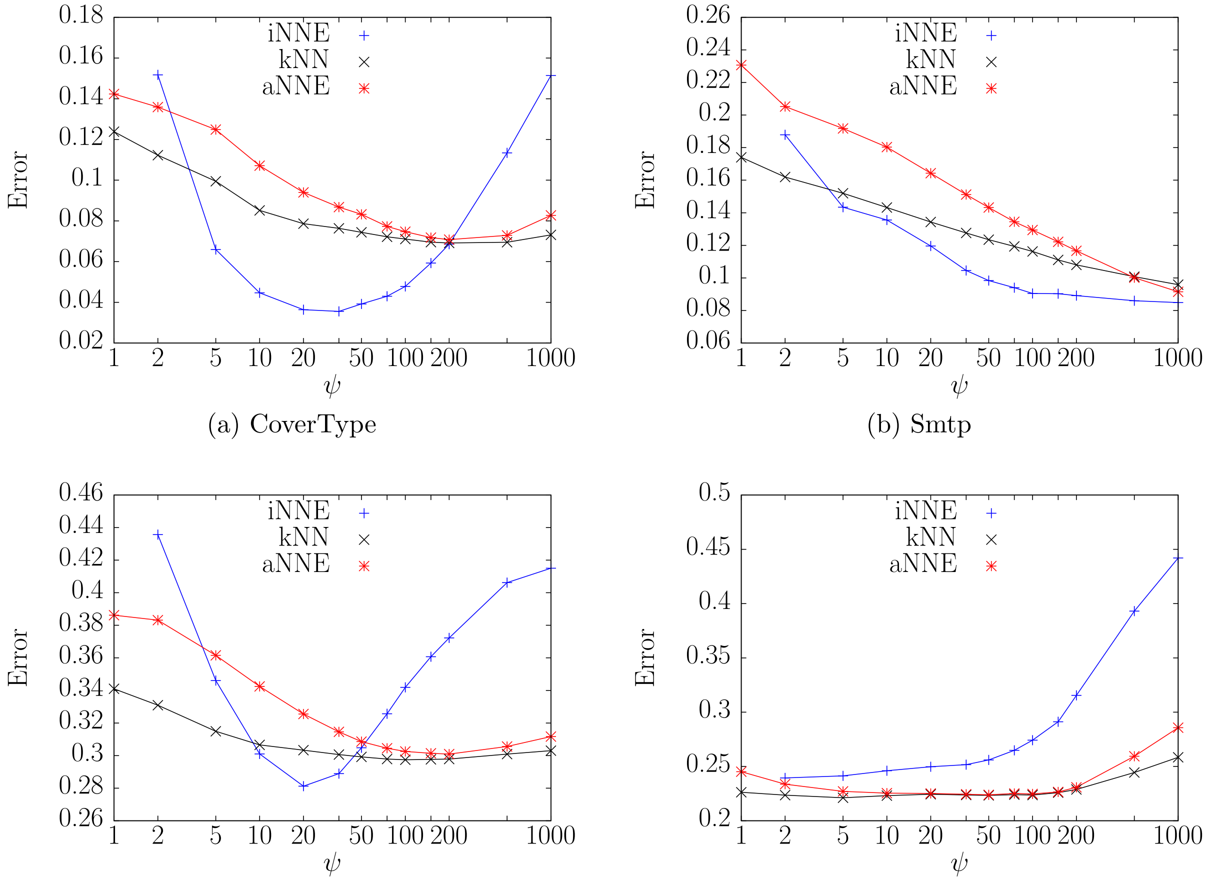
Table 3 summarises the result in Figure 6 by showing the optimal ψ_{opt} for iNNE, aNNE and kNN[4] in each dataset. Recall that $\psi_{opt} < \psi_{max}$ shows the gravity-defiant behaviour. As we have used ψ up to 1000 in the experiment, $\psi_{opt} = \psi_{max} = 1000$ shows the gravity-compliant behaviour. All three anomaly detectors exhibit the gravity-defiant behaviour on all datasets, except the Smtp dataset.

Table 3: ψ_{opt} for iNNE, aNNE and kNN, where $\psi_{max} = 1000$ and $t = 100$.

Data	iNNE	aNNE	kNN
CoverType	35	200	$200t$
Mulcross	2	10	$5t$
Smtp	1000	1000	$1000t$
U2R	20	200	$100t$
P53Mutant	2	20	$75t$
Mammograhpy	200	500	$500t$
Har	2	50	$5t$
ALOI $C = 10$	150	200	$200t$

One interesting result in Table 3 is that ψ_{opt} for iNNE is significantly smaller than those for aNNE and kNN on all datasets. The only exception is the Smtp dataset. Figure 7

[4] Note that the actual optimal training set size for kNN is $t\psi_{opt}$.



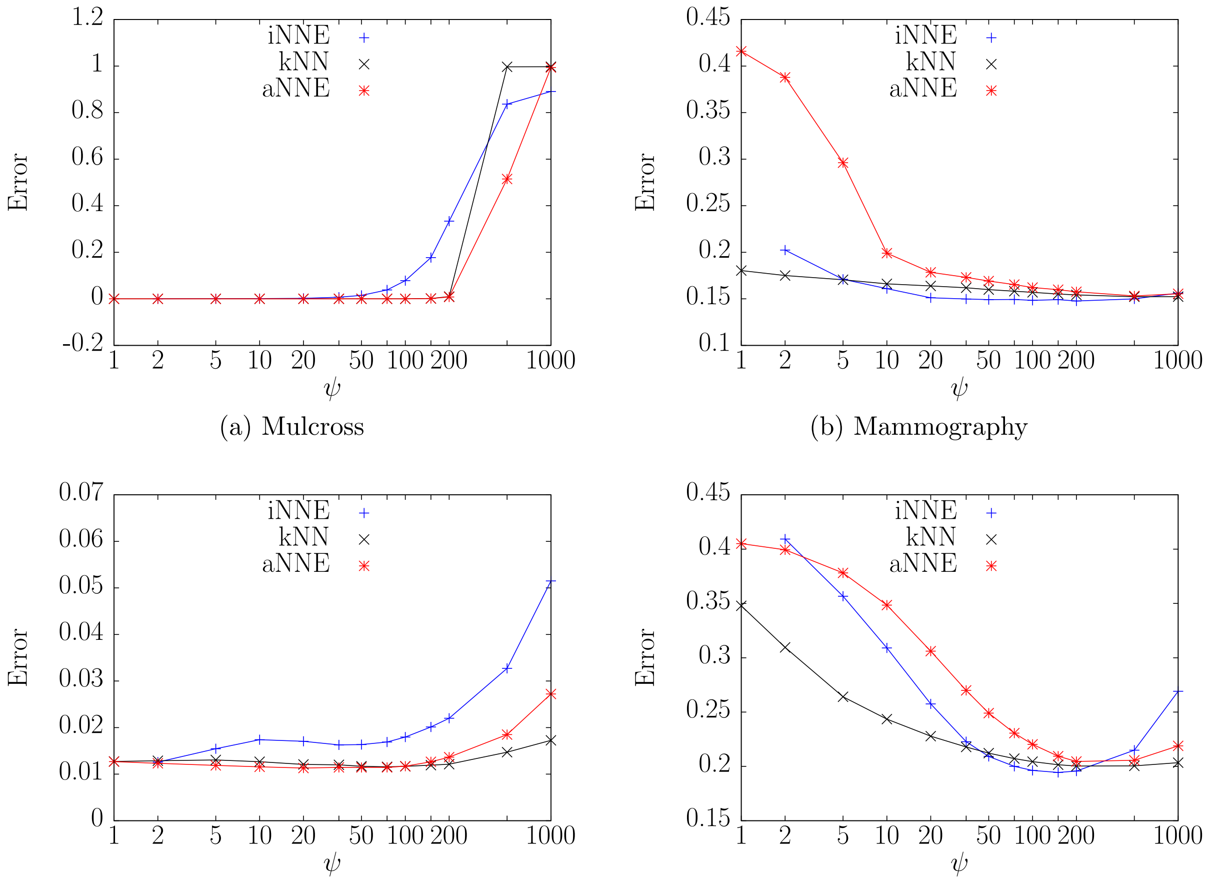
(a) CoverType

(b) Smtp

(c) P53Mutant

(d) Har

Figure 6: Learning curves of iNNE, kNN and aNNE [Section 3.2]: Error is defined as 1-AUC. kNN uses training set size of $t\psi$ and $k = \sqrt{t\psi}$; aNNE and iNNE are using $t = 100$.



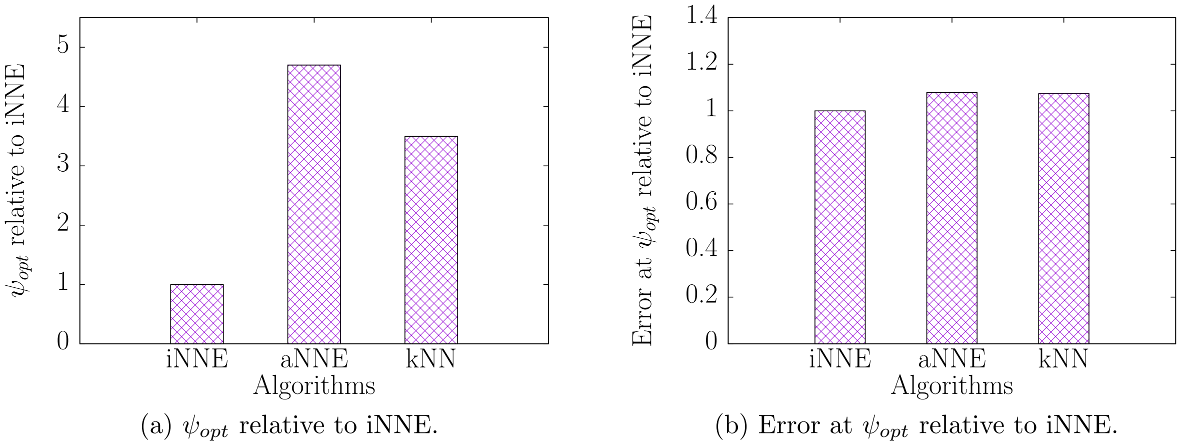
(a) Mulcross

(b) Mammography

(c) U2R

(d) ALOI $C = 10$

Figure 6 (continue) : Learning curves of iNNE, kNN and aNNE: Error is defined as 1-AUC. kNN uses training set size of $t\psi$ and $k = \sqrt{t\psi}$; aNNE and iNNE are using $t = 100$.



(a) ψ_{opt} relative to iNNE.

(b) Error at ψ_{opt} relative to iNNE.

Figure 7: Geometric mean of ψ_{opt} and error at ψ_{opt} relative to iNNE over eight datasets.

shows that the geometric means of ψ_{opt} and error at ψ_{opt} relative to iNNE over the eight datasets. This result shows that aNNE and kNN require about 5 and $4t$ times ψ_{opt} of iNNE, respectively, in order to achieve the optimal detection performance. While all three algorithms have about the same optimal detection performance overall, iNNE has the best on four datasets, equal on two and worse performance than aNNE and kNN on two.

Another interesting observation in Figure 6 is that the learning curves of iNNE almost always have steeper gradient than those of aNNE and kNN.

3.3 Intuition of why small data size can yield the best performing 1NN ensembles

There is no magic to the gravity-defiant algorithms such as aNNE and iNNE which manifest that small data size yields the best performing model. Our result does not imply that less data the better or in the limit zero-data does best. But it does imply that, under some data distribution, it is possible to have a good performing aNNE where each model is trained using one instance only!

We provide an intuitive example as follows. Consider a simple example that all normal instances are generated from a Gaussian distribution. Assume an oracle which provides the representative exemplar(s) of the given dataset for an 1NN anomaly detector. In this case, the only exemplar required is the instance which locates at the centre of the Gaussian distribution. Using the decision rule in Eq (1), where the oracle-picked exemplar is the only instance in $\mathcal{D}$, anomalies are those instances which have the longest distances from the centre, i.e., at the outer fringes of the Gaussian distribution. In this albeit ideal example, $\psi = 1$ for 1NN (as a single model) is sufficient to produce accurate detection. In fact, $\psi > 1$ can yield worse detection accuracy because $\mathcal{D}$ may now contain anomalies when they exist in the given dataset. Both the lower and upper bounds in our theoretical analysis also yield $\psi_{opt} = 1$ for the case of sharp Gaussian distribution.

In practice, we can obtain a result close to this oracle-induced result by random subsampling, as long as the data distribution admits that instances close to the centre has a higher probability of being selected than instances far from the centre, which is the case for sharp Gaussian distribution. Then, an average of an ensemble of 1NN derived from multiple samples $\mathcal{D}_i$ of $\psi = 1$ (randomly selected one instance) will approximate the result achieved by the oracle-picked exemplar. Pang et al. [2] report that an ensemble of 1NN (which is the same as aNNE) achieves the best or close to the best result on many datasets using $\mathcal{D}_i$ of $\psi = 1$!

In a complex distribution (e.g., multiple peaks and asymmetrical shape), the oracle will need to produce more than one exemplar to represent fully the structure of the data distribution in order to yield good detection accuracy. For those distributions with moderate complexity, this number can still be significantly smaller than the size of the given dataset. Pang et al. [2] report that 13 out of the 15 real-world datasets used (having data sizes up to 5 million instances) require $\psi \leq 16$ in their experiments. Note that the dataset size is irrelevant in terms of the number of exemplars required in both the intuitive example and complex distribution scenarios, as long as the dataset contains sufficient exemplars which are likely to be selected to represent the data distribution.

Sugiyama and Borgwardt [9] have previously advocated the use of 1NN (as a single model) with a small sample size and provided a probabilistic explanation which can be paraphrased as follows: a small sample size ensures that the randomly selected instances are likely to come from normal instances only; increasing the sample size increases the chance of including anomalies in the sample which leads to an increased number of false negatives (of predicting anomalies as normal instances).

The above intuitive example and our analyses based on computational geometry further reveal that the geometry of normal instances and anomalies plays one of the key roles in determining the optimal sample size—that signifies the gravity-defiant behaviour of 1NN-based anomaly detectors.

3.4 Which nearest neighbour anomaly detector to use?

A recent investigation [10] has highlighted the difficulty of using kNN because the accuracy of kNN-based anomaly detectors depends on not only the bias-variance trade-off parameter k, but also the data size. Furthermore, the bias-variance trade-off is delicate in kNN because a change in k alters both bias and variance in opposite directions (see Table 4). Our theoretical analysis points to an additional issue, i.e., kNN which insists on using all the available data (as dictated by the conventional wisdom) has no means to reduce the risk of anomaly contamination in the training dataset.

In contrast, our theoretical analysis reveals that, by using 1NN, the risk of anomaly contamination in the training sample can be controlled by selecting an appropriate sample size (ψ). The previous analysis on ensemble of 1NN[5] density estimator [14] shows that the ensemble size (t) can be increased independently to reduce the variance without affecting the bias (see the result shown in Table 4).

Table 4: Squared bias and variance of kNN (for large k) and LiNearN (for $d > 1$), and their bias-variance trade-off parameters. The analytical results are extracted from [13] and [14].

	kNN	LiNearN
Squared bias	$O((k/n)^{\frac{4}{d}})$	$O(\psi^{-2/d})$
Variance	$O(k^{-1})$	$O(t^{-1}\psi^{1-2/d+\epsilon}\Psi^{-1})$
Bias-variance trade-off parameter	k	ψ

where ψ is the sample size used to build the hyperspheres, Ψ is the sample size used to estimate the density in each hyperspheres, t is the ensemble size, d is the number of dimensions, ϵ is a constant between 0 and 1; n is the size of the given dataset.

In addition, our empirical results show that both aNNE and kNN have approximately the same detection accuracy, but kNN requires approximately t times ψ_{opt} of aNNE in order to achieve its optimal detection accuracy[6]. Moreover, searching for ψ (which is usually significantly less than k and does not depend on the data size) is a much easier task than searching for k which is a monotonic function of data size [13]. All in all, we recommend ensembles of 1NN over kNN.

Between the two ensembles of 1NN, we recommend iNNE over aNNE because it reaches

[5] Note that iNNE is a simplified version of LiNearN [14] which does not need a second sample, i.e., Ψ shown in Table 4 is not relevant to iNNE or aNNE.

[6] Similar result applies to ensemble of LOF ($k = 1$) versus LOF.

its optimal detection accuracy with a significantly smaller sample size.

Comparisons with other state-of-the-art anomaly detectors can be found in [7] and [2].

3.5 Implications and potential future work

Both of our theoretical analyses and empirical result reveal that any changes in $\mathcal{X}$ or $\mathcal{X}_N$ lead to changes in nearest neighbour distances. In an unsupervised learning setting, changes in $\mathcal{X}$ or $\mathcal{X}_N$ are usually unknown and difficult to measure in practice. Yet any change that leads to the change in detection error can be measured in terms of nearest neighbour distance: If anomalies' nearest neighbour distances become shorter (or normal instances' nearest neighbour distances become longer), then we know that the detection error has increased, and vice versa. This is despite the fact that the source(s) of the change or the prediction error cannot be measured directly in an unsupervised learning task where labels for instances are not available at all times. However, Δ_A, the average nearest neighbour distance of anomalies, can be easily obtained in practice by examining a small proportion of instances which have the longest distances to their nearest neighbours in the given dataset (so can Δ_N by examining a portion of instances which have the shortest distances to their nearest neighbours), even though the labels are unknown.

This knowledge has a practical impact. In a data stream context, for example, timely model updates are crucial in maintaining the model's detection accuracy along the stream; and the updates rely on the ability to detect changes and the type of change in the stream (for example, whether they are due to changes in anomalies or normal clusters or both.) We are not aware of any good guidance with respect to change detection under different change scenarios in the unsupervised learning context. The majority of current works in data streams focus on supervised learning.

Our finding suggests that the net effect of any of these changes can be measured in terms of nearest neighbour distance, if a nearest neighbour anomaly detector such as iNNE or aNNE is used. This significantly reduces the type and the number of measurements required for change detection. The end result is a simple, adaptive and effective anomaly detector for data streams. A thorough investigation into the application of this finding in data streams will be conducted in the future.

The revelation of the gravity-defiant behaviour of nearest neighbour anomaly detectors invites broader investigation. Do other types of anomaly detectors, or more generally, learning algorithms for other data mining tasks also exhibit the gravity-defiant behaviour? In a complex domain such as natural language processing, millions of additional data has been shown to continue to improve the performance of trained models [12, 11]. Is this the domain for which algorithms always comply with the learning curve? Or there is always a limit of domain complexity, over which the gravity-defiant behaviour will prevail. These are open questions that need to be answered.

3.6 Comparison with conventional ensemble methods

Given the theoretical results, the third aim of this project (i.e., identify the key differences between the gravity defiant ensemble approach and the conventional ensemble approach) becomes irrelevant for the following reasons:

1. The revelation that the nearest neighbour approach to anomaly detection has gravity-defiant behaviour, and an ensemble is not the cause of this behaviour (as initially thought when the project proposal was written). This relinquishes the need for a comparison with existing ensemble methods.

2. Existing ensemble methods such as Bagging and Boosting have never been shown to be successfully in anomaly detection. This is despite the fact that they are very successful in improving the predictive accuracy of classification models such as decision trees and kNN.

It is a worthy investigation in understanding the reason why conventional ensemble methods are gravity-compliance algorithms. But, this is not the main issue in this project—gravity-defiant algorithms. We have thus decided to sidestep the third aim, and focus our attention to deepen our understanding of gravity-defying anomaly detectors. As a result of this focus, we have produced a rigorous theory and empirical evaluation on gravity-defying anomaly detectors.

3.7 The effect of mass-based dissimilarity on aNNE

This section reports the results of our investigation with respect to the fourth project aim: Reveal the effect of new mass-based similarity measures on conventional algorithms and new algorithms in terms of learning curve. We have chosen aNNE in this investigation because it is the focus of our theoretical analyses.

Table 5: ψ_{opt} and error of aNNE for the real data sets. iForest settings for m_e: $\psi = 256$ and $t = 100$. The results are based on a single run of the entire data set. ALOI C=10 with three complexities are used.

				ψ_{opt}		Error	
Dataset	Size n	d	Anomaly Class	ℓ_2	m_e	ℓ_2	m_e
annThyroid	7,200	6	7.42% anomalies	1,000	10	0.2623	0.1908
CoverType	286,048	10	class 4 (0.9%) vs. class 2	500	2	0.0669	0.0945
Har	4728	561	sitting, standing & laying (1.2%)	75	5	0.1957	0.1809
Mammograhpy	11,183	6	class1 (2.32%)	500	1	0.1545	0.1237
mfeat	410	649	2.44% anomalies	10	5	0.0175	0.0105
Mulcross	262,144	4	1% anomalies	75	1	0.0001	0.0001
P53Mutant	31,159	5408	active (0.5%) vs. inactive	150	75	0.2935	0.2422
Smtp	95,156	3	attack (0.03%)	1,000	10	0.0901	0.118
U2R	60,821	34	attack (0.37%)	100	1	0.0119	0.0099
ALOI low		64		150	5	0.1157	0.1981
ALOI medium		64		150	10	0.1974	0.1338
ALOI high		64		1,000	35	0.3694	0.0692

The previous AOARD supported project has revealed that mass-based dissimilarity is a better measure than the commonly used distance measure because the former is data dependent and the latter is data independent. Previous investigations on mass-based dissimilarity [8, 4] was conducted in information retrieval and kNN classification tasks.

Here, we use the latest mass-based dissimilarity [5], which is implemented using iForest [6].

Table 5 shows the results of aNNE using ℓ_2 and mass-based dissimilarity m_e [5].

Two interesting observations are: (a) m_e always produces smaller ψ_{opt} than ℓ_p; and (b) m_e produces lower error than ℓ_p on eight out of the twelve datasets. This shows that the mass-based dissimilarity is a better measure than distance measure in terms of using a small dataset to train a good performing anomaly detector.

This finding invites further investigations to uncover the reason of m_e's superior performance and the condition under which m_e produces higher error than ℓ_p (on three out of the twelve datasets shown in Table 5).

4 Conclusions

The two-year project has exceeded the planned expectation in investigating the gravity-defiant behaviour of nearest neighbour anomaly detectors.

As far as we know, this is the first work which investigates algorithms that defy the gravity of learning curve; and it is also the first time that computational geometry is used to analyse the behaviour of learning algorithms. It provides concrete evidence that there are gravity-defiant algorithms which produce good performing models with small training sets; and models trained with large data sizes perform worse.

The theoretical analysis based on computational geometry gives us an insight into the behaviour of the nearest neighbour anomaly detector. It shows that the AUC changes according to $\psi\alpha^{\psi} \langle\rho\rangle_{\psi}$, influenced by three factors: the proportion of normal instances (α), the radii (ρ) of $\mathcal{X}$ and $\mathcal{X}_N$, and the sample size (ψ) employed by the nearest neighbour-based anomaly detector. Because ψ and $\alpha^{\psi} \langle\rho\rangle_{\psi}$ are monotonic functions changing in opposite directions, an overly large sample size amplifies the negative impact of $\alpha^{\psi} \langle\rho\rangle_{\psi}$, leading to higher error at the tail end of the learning curve—the gravity-defiant behaviour.

We also discover that any change in $\mathcal{X}$ or $\mathcal{X}_N$, which varies the detection error, manifests as a change in nearest neighbour distance such that the detection error and anomalies' nearest neighbour distances change in opposite directions. Because nearest neighbour distance can be measured easily and other indicators of detection error are difficult to measure, it provides a unique useful practical tool to detect change in domains where any such changes are critical in their change detection operations, e.g., in data streams.

The knowledge that some algorithms can achieve high performance with a significantly small sample size is highly valuable in the age of big data because these algorithms consume significantly less computing resources (memory space and time) to achieve the same outcome as those require a large sample size.

In Paper [1] (Section 8), we argue that existing bias-variance analyses on kNN-based density estimators are not an appropriate tool to be used to explain the behaviour of kNN-based anomaly detectors; and the analysis on kNN does not apply to 1NN or ensemble of 1NN on which our analysis targets. In addition, we further uncover that 1NN is not a poor cousin of kNN, rather an ensemble of 1NN has an operational advantage over kNN or an ensemble of kNN: It has only one parameter, i.e., sample size, rather than two parameters, k and data size, that influence the bias—this enables it to have a simpler parameter search. In the age of big data, the most important feature of ensemble of 1NN is that it has significantly smaller optimal sample size than kNN. Unless a compelling reason can be found, we recommend the use of ensemble of 1NN instead of kNN or ensemble of kNN.

Our preliminary investigation using mass-based dissimilarity confirms the results of previous work [8, 4, 5], i.e., mass-based dissimilarity is a better than distance measure. In addition, our investigation further reveals that, in comparison with distance measure, mass-based dissimilarity enables even smaller datasets to be used to train a good performing nearest neighbour anomaly detector.

5 List of Publications and Significant Collaborations that resulted from AOARD supported projects

5.1 List of peer-reviewed journal publications:

[1] Kai Ming Ting, Takashi Washio, Jonathan R. Wells and Sunil Aryal. (2017) Defying the gravity of learning curve: a characteristic of nearest neighbour anomaly detectors. *Machine Learning.* Vol 106, Issue 1, 55-91. doi:10.1007/s10994-016-5586-4.

[2] Guansong Pang, Kai Ming Ting, David Albrecht, Huidong Jin. (2016) ZERO++: Harnessing the power of zero appearances to detect anomalies. *Journal of Artificial Intelligence Research.* Vol 57, 593-620.

[3] Ye Zhu, Kai Ming Ting, Mark James Carman. (2016) Density-ratio based clustering for discovering clusters with varying densities. *Pattern Recognition.* Vol 60, Issue C, 983-997.

[4] Sunil Aryal, Kai Ming Ting, Takashi Washio, Gholamreza Haffari. (2017) Data-dependent dissimilarity measure: an effective alternative to geometric distance measures. *Knowledge and Information Systems.* DOI: 10.1007/s10115-017-1046-0.

5.2 List of peer-reviewed conference publications

[5] Kai Ming Ting, Ye Zhu, Mark James Carman, Yue Zhu, Zhi-Hua Zhou. (2016) Overcoming Key Weaknesses of Distance-based Neighbourhood Methods using a Data Dependent Dissimilarity Measure. *Proceedings of The ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 1205-1214.

5.3 Reference

[6] Fei Tony Liu, Kai Ming Ting, and Zhi-hua Zhou. (2008) Isolation Forest. *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining.* Washington, DC, USA: IEEE Computer Society, 413–422.

[7] Tharindu Bandaragoda, Kai Ming Ting, David Albrecht, Fei Tony Liu, and Jonathan R. Wells. (2014). Efficient anomaly detection by isolation using nearest neighbour ensemble. *Proceedings of the 2014 IEEE international conference on data mining, workshop on incremental classification, concept drift and novelty detection*. 698-705.

[8] Sunil Aryal, Kai Ming Ting, Gholamreza Haffari and Takashi Washio. (2014) mp-dissimilarity: A data dependent dissimilarity measure. *Proceedings of the 2014 IEEE International Conference on Data Mining*. 707-711.

[9] Sugiyama, M., and Borgwardt, K. (2013). Rapid distance-based outlier detection via sampling. *Advances in Neural Information Processing Systems*. 26, 467–475.

[10] Aggarwal, C. C., and Sathe, S. (2015). Theoretical foundations and algorithms for outlier ensembles. *SIGKDD Explorations*. 17(1), 24–47.

[11] Banko, M., and Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. *Proceedings of the 39th annual meeting on association for computational linguistics, association for computational linguistics*. 26–33.

[12] Halevy, A., Norvig, P., and Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*. 24(2), 8–12.

[13] Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd ed.). San Diego: Academic Press.

[14] Jonathan R.Wells, Kai Ming Ting, and Takashi Washio (2014). LiNearN: A new approach to nearest neighbour density estimator. *Pattern Recognition*. 47(8), 2702–2720.

5.4 Significant collaborations

Takashi Washio of Osaka University has contributed significantly in this project, resulting in two joint papers [1, 4]. Monash University colleagues, Mark Carman, Gholamreza Haffari and David Albrecht, have collaborated in this project; and they are the co-authors in four papers.

Papers produced as a result of previous AOARD projects

- Paper [5] provides the first generic version of mass-based similarity measure and a tree implementation.
- Paper [2] presents a new unsupervised technique to use zero appearance in subspaces to detect anomalies.
- Paper [3] uncovers the exact condition under which density-based clustering such as DBSCAN fails; and introduces three principled ways to overcome this shortcoming. This has informed the work reported in Paper [5].
- Paper [4] is the journal version of the ICDM2014 paper on mass-based dissimilarity [8]. It provides an efficient implementation and expands the investigation into measures which can deal with categorical attributes, numeric attributes and mixed attributes.

Software Download

The source code of mass-based dissimilarity [5] can be obtained at `https://sourceforge.net/projects/mass-based-dissimilarity/`.